Svetina and Rutkowski *Large-scale Assessments in Education* 2014, **2**:4
http://www.largescaleassessmentsineducation.com/content/2/1/4

Large-scale Assessments in Education
a SpringerOpen Journal

RESEARCH Open Access

Detecting differential item functioning using generalized logistic regression in the context of large-scale assessments

Dubravka Svetina* and Leslie Rutkowski

* Correspondence:
dsvetina@indiana.edu
Department of Counseling and Educational Psychology, Indiana University, Bloomington, USA

Abstract

Background: When studying student performance across different countries or cultures, an important aspect for comparisons is that of score comparability. In other words, it is imperative that the latent variable (i.e., construct of interest) is understood and measured equivalently across all participating groups or countries, if our inferences regarding performance can be regarded as valid. Relatively fewer studies examined an item-level approach to measurement equivalence, particularly in settings where a large number of groups is included.

Methods: This simulation study examines item-level differential item functioning (DIF) in the context of international large-scale assessment (ILSA) using a generalized logistic regression approach. Manipulated factors included the number of groups (10 or 20), magnitude of DIF, percent of DIF items, the nature of DIF, as well as the percent of affected groups with DIF.

Results: Results suggested that the number of groups did not have an effect of the performance of the method (high power and low Type I error rates); however, other factors had impacted the accuracy. Specifically, Type I error rates were inflated in non-DIF conditions, while they were very conservative in all of the DIF conditions. Power was generally high, in particular in conditions where DIF magnitude was large, with one exception – in conditions where DIF was introduced in difficulty parameters and the percent of DIF items was 60.

Conclusions: Our findings presented a mixed picture with respect to the performance of the generalized logistic regression method in the context of large number of groups with large sample sizes. In the presence of DIF, the method was successful in distinguishing between DIF and non-DIF, as evidenced by low Type I error and high power rates. On the other hand, however, in the absence of DIF, the method yielded increased Type I errors.

Background

When studying student performance across different countries or cultures, an important aspect for comparisons is that of score comparability. In other words, it is imperative that the latent variable (i.e., construct of interest) is understood and measured equivalently across all participating groups or countries, if our inferences regarding performance can be regarded as valid. The psychometric property that typically must hold for scores to be comparable is known as *measurement invariance* (Meredith 1993),

absence of *differential item functioning* (Hambleton et al. 1991; Mellenbergh 1994; Swaminathan & Rogers 1990), or *lack of bias* (Lord 1980). Regardless of the term used, the literature on scale score equivalence in large-scale achievement tests has received considerable attention (e.g., Ercikan 2002; Hambleton 2002; Oliveri et al. 2012). Many of these investigations have focused on pairwise comparisons of countries (Oliveri et al. 2012; Oliveri 2012), the latter of which uses both empirical and simulated data. In the context of international attitude and behavior scale development, researchers have also used multiple-group confirmatory factor analysis (MG-CFA, Jöreskog 1971; OECD 2010). Under an MG-CFA framework, researchers evaluate hypotheses of configural, metric, and scalar invariance and conduct difference tests to evaluate which level of invariance is supported by the responses.

Relatively fewer studies examined an item-level approach to measurement equivalence, particularly in settings where a large number of groups is included. One prominent exception includes the Programme for International Student Assessment (PISA), which uses an ANOVA-like approach to examine item-by-country interactions (OECD 2012). This general paucity might stem from the fact that, until recently, methodologies that would allow for simultaneous examination of a large number of groups for comparisons were not available. As we note below, generalized versions of some methodologies used for the two group comparisons are now available to researchers for conducting comparisons across more than two groups; however, little information on the performance of these methods exists, in particular in the context of large-scale studies such as PISA or the Trends in International Mathematics and Science Study (TIMSS). Also, to our knowledge, relatively little literature exists that questions the current practices of measurement invariance via MG-CFA framework in such settings.

A recent study by Rutkowski and Svetina (2014) provides some evidence that typically recommended criteria for evaluating invariance in MG-CFA (e.g., change in fit statistics; chi-square difference tests) may not always be appropriate when large numbers of groups are compared, as is typically the case in an international assessment or survey context. More specifically, Rutkowski and Svetina studied the performance of MG-CFA and associated fit criteria when the number of groups are relatively large and found that as a measure of overall model fit, the chi-square statistic was not useful, as it suggested strong model-data misfit across all studied conditions. Furthermore, the chi-square difference test (as typically applied) was also too conservative in the studied context. While results for the chi-square test were not surprising, fit indices results were unexpected to some extent. Namely, in examining the overall fit indices (RMSEA, CFI, TLI, and SRMR), the authors found that currently accepted cutoffs for the CFI, TLI, and SRMR as an overall fit indicators were generally suitable (although SRMR in isolation produced somewhat conservative results). However, when considering the relative fit (i.e., change in a fit index from configural to metric or metric to scalar) via changes in CFI and RMSEA, the authors found that the change in RMSEA associated with the metric invariance hypothesis to be increased or larger than a typically accepted difference of .010.

As mentioned above, there has been little research in DIF settings for multiple groups scenarios despite large interest in cross-cultural and multilingual research (Fidalgo & Scalon 2010). This may be partly due to the availability of methods that allow for more

than two group comparisons simultaneously, given that a more common alternative approach is somewhat tedious and it requires several steps; namely, it involves pairwise comparisons among all group pairs. In practice, this pairwise comparisons method has been adopted by several researchers, and several variations on this approach exist. For example, as outlined in Ellis and Mead (2000), one way to conduct pairwise comparisons and multiple group DIF analysis is to consider in the first step fitting a separate item response theory (IRT) model to obtain item parameters for each group or country (of course, stringent IRT assumptions are assumed here to have been met prior to model fitting). Choice of software may depend on the analyst's skill or software availability, but typical choices include BILOG-MG (Zimowski et al. 1996) or MULTILOG 7.03 (Thissen et al. 2003). Then, an analyst would link the parameters using a linking software, for example, EQUATE 2.1 (Baker 1993). After a successful linking procedure, item parameters would be compared among the groups.

Versions of this approach have been adopted in literature on cross-cultural phenomena, including translation of surveys measuring personality (e.g., Ellis & Mead 2000), global employee attitudes (e.g., Ryan et al. 2000), dominance (e.g., Kulas et al. 2011), emotional functioning (e.g., Petersen et al. 2003), health/quality of life (e.g., Scott et al. 2006 2007), and reading (e.g., Glas & Jehangir 2013; Oliveri & von Davier 2011, 2014), to name a few. Specifically, Oliveri and von Davier (2011, 2014) have provided empirical evidence that allowing a subset of item parameters to be uniquely estimated offers one way to improve model-to-data fit and reduces problems with comparability across heterogeneous populations and associated parameter estimate bias. Additionally, Glas and Jehangir (2013) proposed using a Lagrange multiplier (LM) test to identify poorly fitting (polytomous) background items where one country serves as the reference country and all other countries are pooled together to serve as the focal country. The single worst fitting item is identified and allowed to be freely estimated; the process is repeated until 95% of the residuals are sufficiently low or four items have been freely estimated for any given country. This approach is to some extent similar to the Oliveri and von Davier method in that some items are freely estimated, although Glas and Jehangir considered background questionnaire items and the process is iterative and, possibly, time intensive.

In the current study, emphasis is given to the investigation at an item-level, rather than a scale level but where achievement (rather than background) items are estimated simultaneously rather than using a multi-step process adopted in the above mentioned studies. The emphasis here on item-level is given based on the following reasons. First, as noted above, typical cut-off criteria within the framework of MG-CFA may not be appropriate for comparing large numbers of groups, hence the item-level analysis may prove more suitable. Second, motivated by invariance research in language testing and translations, Zumbo (2003) investigated whether item-level DIF manifested itself in the scale-level analysis. Findings indicated that limiting investigations to the scale-level only is often not sufficient since item-level DIF may be obscured in a scale-level analyses (p. 146). In other words, item-level DIF should also be examined in conjunction with scale-level investigations because additional insights may be gained. As such, the current paper investigates the performance of one method of DIF detection – generalized logistic regression – under conditions that closely follow international assessments.

The current research attempts to contribute to the literature on DIF when a large number of groups is considered and the analyses are conducted at an item-level. Specifically, using a Monte Carlo study, we are interested in answering the following research question: How does a generalized linear logistic regression method perform in identifying DIF when a large number of groups is considered? The remainder of the paper is organized as follows. In the next section, we introduce methods appropriate for item-level DIF analysis, with a particular focus on the approach used in the current study – the generalized logistic regression method. Next, we describe the methods and provide a rationale for the choices of the study design. Results, presented separately for 10- and 20-group conditions, follow. Lastly, we conclude with a summary and a discussion of limitations.

Methods to examine item-level DIF

Several methods exist to investigate item-level DIF, including the Mantel-Haenszel approach (Holland & Thayer 1988), logistic regression (Swaminathan & Rogers 1990), which conditions on the observed scores, or methods rooted in item response theory (IRT), where item characteristic functions are estimated for each group and then compared to investigate DIF, such as Lord's chi-square test (Lord 1980) or Raju's method (Raju 1988, 1990).

Several factors likely influence any analysts choice of DIF method, including a preference for one of the aforementioned frameworks or methodological approaches (IRT vs. non-IRT), the type of DIF effect of interest (uniform, nonuniform, or both), underlying assumptions (parametric or nonparametric), the number of groups under comparison (two or more), and data characteristics (e.g., dichotomous or polytomous scoring; missing data). Over the last several decades, the literature has featured many DIF methods suitable for two groups (one *reference* and one *focal*) and reviews of the methods suitable for dichotomous or polytomous items (e.g., Camilli & Shepard 1994; Millsap & Everson 1993; Penfield & Lam 2000; Potenza & Dorans 1995). Furthermore, performance of different DIF methods have been examined (e.g., Li et al. 2012; Penfield 2001), including those that allow for multiple groups (i.e., > 2), such as generalized Mantel-Haenszel (e.g., Fidalgo & Madeira 2008; Fidalgo & Scalon 2010; Penfield 2001), generalized Lord's test (e.g., Kim et al. 1995) and generalized logistic regression (Magis et al. 2013; Magis et al. 2011).

Given our study design, discussed subsequently, and our goal of studying item-level DIF when the number of groups compared is greater than two, we consider the performance of the generalized logistic regression method (Magis et al. 2011). The method is an extension of the well-known logistic regression approach to investigate DIF between two groups, as proposed by Swaminathan and Rogers (1990). Our choice of the generalized logistic regression as the method of choice to study DIF was based on the following. As suggested by Magis et al., the method allows for simultaneous estimation of group parameters, hence it avoids multistep, pairwise comparisons process (and it eliminates the necessity to use multiple software packages). In addition, its flexible framework allows for investigation of uniform, nonuniform, or both types of DIF within subgroups, which is an attractive feature since in international settings, it is likely that either type of DIF is plausible, given the inherent diversity and complexity of the studied context. And more so, it is an improvement

over other generalized methods, such as the generalized Mantel-Haenszel method, which only tests for uniform DIF. Lastly, no merging of the focal groups is necessary.

As noted in the Methods section below, for our analysis, we used the generalized logistic regression approach implemented as a function in difR package (Magis et al. 2013) in R (R Development Core Team 2012). It is an appropriate method when data are dichotomously scored and where more than two groups are considered. Furthermore, this method allows for studying both uniform and nonuniform DIF and it investigates one item at the time, such that in the process, the remaining items not under consideration are assumed to be non-DIF[a]. As the method of analysis in this study, we briefly highlight the main components of the generalized logistic regression. Interested readers are directed to the original work by Magis et al. (2011) for complete details.

The generalized logistic regression DIF model, as presented by Magis et al. (2011), has the following form:

$$\text{logit}(\pi_{ig}) = \alpha + \beta S_i + \alpha_g + \beta_g S_i, \tag{1}$$

where π_{ig} is the probability of examinee i from group g correctly responding to an item, logit is the natural log of the odds of correctly answering an item, α and β are common intercept and slope parameters (i.e., for all groups), α_g and β_g are group-specific slope and intercept parameters, and S_i is the total test score for examinee i, which serves as a matching variable and a proxy for the ability level of the examinee. For model identification purposes, group-specific parameters for the reference group (denoted as $g = 0$), α_0 and β_0, are set to zero. In other words, if $g = 0$ (i.e., reference group), $\text{logit}(\pi_{ig}) = \alpha + \beta S_i$, and if $g \neq 0$ (i.e., focal groups, $g = 1, 2, \ldots, F$), $\text{logit}(\pi_{ig}) = (\alpha + \alpha_g) + (\beta + \beta_g)S_i$. An item is said to contain DIF if the probability π_{ig} varies across the groups of examinees (i.e., there is an interaction between the group membership and the item response). This occurs when at least one of the group parameters, α_g or β_g, is different from zero. If all group-specific parameters equal zero, an analyst would conclude that no DIF is present.

According to Magis et al. (2011), three types of DIF can be investigated using this framework: a) uniform DIF (*UDIF*), b) nonuniform DIF (*NUDIF*), and both types of DIF effects together (*DIF*). Tested null hypotheses for these three types include:

$$H_0 : \alpha_1 = \ldots = \alpha_F = \beta_1 = \ldots = \beta_F = 0 \qquad DIF \tag{2}$$

$$H_0 : \beta_1 = \ldots = \beta_F = 0 \qquad NUDIF \tag{3}$$

$$H_0 : \alpha_1 = \ldots = \alpha_F \big| \beta_1 = \ldots = \beta_F = 0 \qquad UDIF \tag{4}$$

Statistically, DIF is assessed by an examination of the model parameters. Let $\boldsymbol{\tau}$ represent a vector of model parameters, $\boldsymbol{\tau} = (\alpha, \alpha_1 \ldots \alpha_F, \beta, \beta_1 \ldots \beta_F)^T$ for *DIF* and *NUDIF* and $\boldsymbol{\tau} = (\alpha, \alpha_1 \ldots \alpha_F, \beta)^T$ for the *UDIF*, and $\hat{\boldsymbol{\tau}}$ be the maximum likelihood estimator of $\boldsymbol{\tau}$. Using maximum likelihood, we can test the null hypotheses in (2) to (4) by different methods, such as the Wald test or the likelihood ratio test (LRT). In the current study,

we utilized LRT as a criterion to detect DIF, where the null (M_0) and alternative (M_1) models were given by:

$$M_0 \equiv logit\left(\pi_{ig}\right) = \begin{cases} \alpha + \beta S_i + \alpha_g & \text{in NUDIF} \\ \alpha + \beta S_i & \text{in DIF and UDIF} \end{cases} \tag{5}$$

and

$$M_1 \equiv logit\left(\pi_{ig}\right) = \begin{cases} \alpha + \beta S_i + \alpha_g + \beta_g S_i & \text{in DIF and NUDIF} \\ \alpha + \beta S_i + \alpha_g & \text{in UDIF} \end{cases} \tag{6}$$

DIF is then tested by the lambda statistic (Wilks 1938), where $\mathbf{\Delta} = -2\log\left(\frac{L_0}{L_1}\right)$, L_0 and L_1 are the corresponding maximum of the likelihoods for the M_0 and M_1, respectively, which follows an asymptotic chi-square distribution with degrees of freedom of the asymptotic null distribution of $\mathbf{\Delta}$ equal to $2F$ for DIF and F for both the *UDIF* and *NUDIF*. In other words, an item would be flagged as DIF when the lambda statistic is sufficiently large, or stated alternatively, when the alternative model is preferred.

Methods

Study design

As stated above, our study is situated within the context of ILSAs; hence, we designed our simulation study as follows. First, we follow the practice of test design and administration by adopting a rotated booklet design (see Table 1). In operational testing, such as in the TIMSS 2007 assessment in grade 4, 10 to 14 items per block were administered and each student received a total of four blocks (two for mathematics and two for science). In our booklet design, we aimed to achieve a similar set up for only one content assessment.

As Table 1 shows, for purposes of our study, each booklet contained three blocks, where each simulee received a single booklet. An approximately equal number of simulees received any one of the seven booklets within a group (with random assignment), and each group was administered all seven booklets. Each block contained 15 items, resulting in 45 items per simulee. The remaining items (those in blocks not administered to any one simulee at the time) were treated as missing by design in the analysis.

Table 1 Block and booklet design for the study for one group

	Block (15 items per block)						
Booklet	**1**	**2**	**3**	**4**	**5**	**6**	**7**
A	x	x		x			
B		x	x		x		
C			x	x		x	
D				x	x		x
E	x				x	x	
F		x				x	x
G	x		x				x

Note. Any one booklet contained a total of 45 items and a total of 105 items were simulated for each group (15 items per block × 7 booklets).

Our study design choices were motivated by what is typically observed in practice with ILSAs. To examine the impact of various factors on DIF detection, we manipulated several variables (see Manipulated factors). To generate the data, we used empirical approximations of item and person parameters from an existing dataset (see Data generation). In what follows, we offer rationale for the choices made in the study design, including manipulated factors (and respective levels) of the simulation study and the process of the selection/modification of item and person parameters.

Manipulated factors

Several factors were manipulated in the study, including

a) number of groups (10 or 20);
b) percent of groups affected by DIF (40% or 70%)
c) percent of DIF items (20%, 40%, or 60%);
d) nature of DIF (difficulty or discrimination); and
e) magnitude of DIF (small or large).

The number of groups examined in the study was set at 10 or 20. One of our goals was to examine the performance of generalized logistic regression model as implemented in the difR package (Magis et al. 2013) in a relatively large group setting; these group sizes approximate more closely the operational context of large-scale surveys. For example, the Teaching and Learning International Survey typically has 20–30 participating educational systems, depending on the grade level (OECD 2010). We do, however, recognize that other assessments, such as PISA or TIMSS, may have 60 participating educational systems or more. One of the primary reasons we focused on what might be considered lower bounds of the number of groups is the computing time required per analysis. In the preliminary analysis of conditions with 30 and 60 groups (with other design choices remaining the same), analysis of one replication required over 3 and 5.5 hours respectively on an IBM e1350 high-performance computing system[b]. We further recognize that the computing time is not necessarily an issue in reality when only one dataset is considered. However, for the purposes of our simulation study, with multiple replications within any one condition, it was necessary to choose a smaller number of groups.

The number of groups affected by DIF was set to be 40% or 70% of total groups. This meant that in conditions with 10 groups, either 4 or 7 groups contained some items with DIF, and in conditions with 20 groups, 8 or 14 groups were affected by DIF. For all conditions, the same arbitrarily selected *reference* group was used, while the remaining groups were treated as *focal*. The assignment of groups affected by DIF was made randomly. The reference group had a moderate sample size ($N = 6370$ in 10-group and $N = 5341$ in 20-group conditions), was the same reference group across all conditions, and its non-DIF item parameters were similar to those of focal groups.

The percent of items that were modeled as DIF-items was set to be 20%, 40%, or 60%, which resulted in 21, 42, or 63 DIF items for selected groups (a total of 105 items were administered for any one group – see more detail below on how these parameters were selected and manipulated). The assignment of DIF items was also drawn at random, and was kept consistent across groups and levels of DIF (i.e., nature of DIF). Our

choice for selecting a wide percent of items to be affected by DIF was primarily motivated by research on large-scale assessments in language or cross-cultural studies, where it is not uncommon to see a large number (percent) of items flagged as DIF-items. For example, Stubbe (2011) investigated DIF in the German-language versions of an international reading assessment, where he found an overall percentage of significant DIF items to be over 74%; percent of DIF items ranged from 71.4% to 77.4% for multiple-choice and constructed responses item types, respectively (2011). Further, Dorans and Middleton (2012) cogently argue, using well-established literature and empirical support, that *language is a condition of measurement* and as such, typical assumptions that identical items translated into testing languages are not necessarily equivalent. These "extreme assumptions" underpinning what the authors term "presumed" linking (in contrast to empirical linking) should be subjected to falsifiable tests as a means of building support for a presumed linking across potentially non-equivalent groups. The authors argue that invariance of relationships among test forms should be examined. Further, necessary conditions for score-scale comparability include a set of equivalent anchor items. As such, our design, even at the extreme, includes a set of non-invariant as well as invariant "anchor" items.

The nature of DIF was examined by changing the value of the difficulty or discrimination parameters for DIF-items in affected groups. This is akin to the investigation of uniform and nonuniform DIF, respectively. The generalized logistic method used in the analysis allows for investigation of both types of DIF; hence, in our analysis, we opted for DIF investigation of both types of DIF across all conditions. Focal groups that did not contain DIF items had difficulty and discrimination parameters equal to those of the reference group.

The magnitude of DIF was examined at two different levels. In conditions with *small* DIF, values for difficulty parameters for affected items differed from non-DIF items by .50 and discriminations varied by .40. In conditions with *large* DIF, differences were 1.00 and .80 for difficulty and discrimination parameters, respectively. Across all of the conditions, differences in value of item parameters was made at random, either in favor of the reference or focal group in question (i.e., differences in parameter values were either added or subtracted from the non-DIF item in question by a random draw). A note should be made regarding our choice of the difference values in introducing DIF. Within typical DIF investigations, some authors suggest that differences of .50 in difficulty, for example, might be considered large DIF (c.f., Goodman et al. 2011). In our study, we used descriptors of *small* and *large* only to differentiate different levels of DIF, not necessarily the classification of the DIF magnitude. More importantly, we wanted to use values that can be found in examining difference in item parameters in large-scale assessments. Specifically, in a study of 21 countries in PISA 2009, Rutkowski and Rutkowski (2014) found that differences in estimates of item difficulties between any two countries could be as large as 6.22 and as small as 1.95, with an absolute average of difficulty difference of .90 and absolute maximum mean difference across all items of 2.96[c] (Rutkowski, L., & Rutkowski, D. One size does not fit all: The impact of item parameter differences and estimation methods on cross-cultural achievement comparisons, submitted manuscript). Similarly, the authors found that the discrimination differences for the same data were found to be in range of .55 and 2.19, respectively, with an average maximum difference of 1.32.

Data generation

In order to simulate our data, we used existing released item parameters that could be found on typical ILSAs. Specifically, for non-DIF conditions, we estimated item parameters from the 2007 TIMSS grade 4 assessment. On 2007 TIMSS, a total of 179 items were released for grade 4 across all content areas, of which 174 were scored dichotomously (i.e., either multiple-choice or constructed response items with only two categories; Olson et al. 2008). From this pool of 174 released items (i.e., their parameters), we randomly drew without replacement a total of 105 items. The 105 items represented the total number of items used in our study, where each of the seven blocks contained 15 items.

Once the 105 item parameters were selected at random, they were treated as non-DIF and were assigned to non-overlapping blocks at random for the reference group and any focal group(s) selected to be non-DIF. Table 2 shows the average, minimum, and maximum value of the difficulty and discrimination item parameters for non-DIF items across the seven blocks. For example, as shown in Table 2, item 1 had an average difficulty of .27 across the seven blocks, and its difficulty parameter ranged from -.23 to 1.06. As Table 2 suggests, item parameters drawn from the released 2007 TIMSS grade 4 assessment varied in both difficulty and discrimination, although these values would be considered typical in such assessments.

Person parameters for each group were drawn randomly from a normal distribution with some mean and standard deviation (see Table 3). The means and standard deviations varied across groups; the differences in population means (standard deviations) were purposefully introduced to resemble parameters likely to be found in large-scale assessments. Values found in Table 3 were obtained by random sampling with replacement of the "standardized" plausible values from 44 participating countries from grade 4 mathematics on TIMSS 2007. Note that for 10-group conditions, every other value for 20-group conditions was used. Although these values are based on plausible values provided by TIMSS, they should not be used as estimate of countries' abilities. Rather, we use them here only as an approximations in order to introduce group differences.

Table 3 also shows that the sample size varied across the groups. These sample sizes were chosen to resemble current large-scale assessments, where group sample sizes are typically large and vary per participating country. Values for sample sizes were randomly drawn from $\sim N(5000, 1000)$ for 20 groups; every other sample size from the 20-group pool was assigned to the 10-group. Within each group, approximately the same number of simulees received each of the booklets, also assigned at random.

We simulated binary responses to items according to a 2-parameter logistic (2-PL) model via the *sim* function implemented in irtoys package (Partchev 2012) in R. The 2-PL specifies the probability of person i endorsing an item j ($X_{ij} = 1$) as:

$$P\left(X_{ij} = 1|\theta_i, a_j, b_j\right) = \frac{e^{a_j\left(\theta_i - b_j\right)}}{1 + e^{a_j\left(\theta_i - b_j\right)}}, \tag{7}$$

where θ_i is the person parameter for person i, b_j is the location parameter (difficulty) for item j, and a_j is the discrimination parameter for item j. A fully factorial design yielded a total of 50 conditions; 2 (number of groups) × 2 (percentage of DIF-affected groups) × 3 (percent of DIF items) × 2 (nature of DIF) × 2 (magnitude of DIF) + 2 (non-DIF conditions), each replicated 100 times.

Table 2 Item difficulty and discrimination parameters used for generation of non-DIF items

Item	$\bar{b}$	b_{min}	b_{max}	$\bar{a}$	a_{min}	a_{max}
Item 1	.27	-.23	1.06	1.00	.57	1.33
Item 2	-.40	−1.58	.35	.97	.66	1.28
Item 3	.31	-.49	.85	.92	.50	1.17
Item 4	-.14	-.94	.45	1.04	.74	1.56
Item 5	.24	-.44	1.04	1.00	.73	1.19
Item 6	.08	-.63	.92	1.06	.80	1.73
Item 7	.12	−1.21	.93	.97	.70	1.63
Item 8	.06	−1.37	.97	.98	.78	1.23
Item 9	.03	-.55	.95	1.08	.55	1.76
Item 10	-.28	−1.23	.73	1.05	.92	1.18
Item 11	-.23	−2.00	1.09	.97	.65	1.23
Item 12	.09	−1.21	.68	.83	.46	1.23
Item 13	.22	-.46	.79	.82	.46	1.39
Item 14	-.04	-.62	.64	1.03	.76	1.55
Item 15	-.35	−1.38	.76	.92	.51	1.31

Note. These are the averages of item parameters across the seven blocks within a country that contains only non-DIF items.

Table 3 Descriptive statistics for group proficiency parameters

Group	*M*	*SD*	*N*	*M*	*SD*	*N*
1	.32	.56	5460	.32	.56	5460
2	.26	.55	6328	−1.15	.72	5733
3	.25	.58	5481	.26	.55	6328
4	.86	.56	5236	.79	.63	5474
5	.29	.69	5362	.25	.58	5481
6	-.30	.69	5537	.26	.64	4179
7	.40	.58	5362	.86	.56	5236
8	-.24	.74	6370	.52	.71	5341
9	−1.47	.74	6545	.29	.69	5362
10	.11	.58	4494	−1.34	.81	5117
11				-.30	.69	5537
12				-.80	.71	5187
13				.40	.58	5362
14				-.30	.69	3983
15				-.24	.74	6370
16				.45	.62	5845
17				−1.47	.74	6545
18				.60	.68	3297
19				.11	.58	4494
20				.50	.51	4571

Note. The means (*M*) and standard deviations (*SDs*) here represent distributional parameters from which *N* samples for each group were drawn.

Analysis

Generated data were analyzed within the framework of generalized logistic regression as implemented in the difR package under *genDichoDif* (method="genLogistic") function (Magis et al. 2013) in R (R Development Core Team 2012). The reference group (group 8) was used across all conditions in both 10- and 20-group cases, while remaining groups were treated as focal[d]. Default options were used for the analysis, including the test of both uniform and non-uniform DIF, LRT was used as the criterion to detect DIF items, and the method was not used iteratively to purify the set of anchor items. Items not administered per any one record were treated as missing in the analysis, although no missingness was introduced in the group membership (i.e., every simulee had a group membership designation).

Performance criteria

In order to evaluate the performance of the generalized logistic regression, as implemented in difR, we report Type I error and power rates. Type I error rates are reported as the proportion of items within a replication that were generated as non-DIF items but which are flagged as DIF in the analysis. Similarly, power rates are computed as the proportion of items that are simulated as DIF items that are also flagged as having DIF. Rates are averaged across 100 replications for each condition separately.

Results

For the purpose of organization, we present our results for the 10-group conditions first, followed by the 20-group conditions. Type I error rate for the NONDIF condition (where data for all 10 groups was modeled using the same set of item parameters) was inflated across the 100 replications; the mean (standard deviation) for the NONDIF condition was .20 (.02), which deviates from a typically used .05 or .01 levels. For the 10-group DIF conditions, regardless of the level or nature of DIF introduced, the Type I error rates were all .00. Given the consistency across all studied conditions, we only mention the Type I error rates of zero and do not present these results in a table. Our finding, with respect to Type I error rates, suggests that the generalized logistic method was accurate in all DIF conditions by not flagging non-DIF items as having DIF.

Table 4 shows the average power rates (and standard deviations) for DIF conditions for the 10-group cases. Rates were generally moderate to high across the studied conditions, with a few exceptions. The highest power rates were noted in conditions where the difference in item parameters for both the discrimination and difficulty were large. This is not unexpected, but it should also be noted that this difference was mostly noted in conditions where discrimination parameters varied (panel (a)); as panel (b) showed, uniform DIF was quite constant across different DIF levels. It was also noted that, generally, the power rates were higher in conditions where the percent of groups affected by DIF was at 40% compared to 70%. This seemed to be the case for DIF conditions where either discrimination or difficulty was invariant.

Several patterns were also noted. For example, in panel (a), where DIF was introduced in the discrimination parameter, the percent of DIF items within an affected group did not exhibit a clear pattern. For example, when discrimination DIF was .40

Table 4 Average power rates across studied conditions for 10 and 20 groups

Panel (a) Invariance in Discrimination Item Parameters									
		Difference in Parameters for DIF Items							
	% of DIF items	**Δ_a = .40% of Groups with DIF Items**				**Δ_a = .80% of Groups with DIF Items**			
		40	**70**	**40**	**70**	**40**	**70**	**40**	**70**
		10 Groups		**20 Groups**		**10 Groups**		**20 Groups**	
Discrimination	20	.70	.51	.70	.55	1.00^{+}	.99	1.00	1.00
		(.08)	(.07)	(.07)	(.07)	(.01)	(.02)	(.00)	(.01)
	40	.65	.57	.64	.61	.99	.98	.99	.98
		(.05)	(.04)	(.03)	(.04)	(.01)	(.01)	(.01)	(.01)
	60	.71	.63	.72	.68	.97	.97	.97	.97
		(.04)	(.04)	(.03)	(.03)	(–)	(.01)	(.01)	(.01)
Panel (b) Invariance in Difficulty Item Parameters									
		Difference in Parameters for DIF Items							
	% of DIF items	**Δ_b = .50% of Groups with DIF Items**				**Δ_b = 1.00% of Groups with DIF Items**			
		40	**70**	**40**	**70**	**40**	**70**	**40**	**70**
		10 Groups		**20 Groups**		**10 Groups**		**20 Groups**	
Difficulty	20	.97	.92	.95	.95	1.00	1.00	1.00	1.00
		(.03)	(.03)	(.01)	(.03)	(.00)	(.00)	(.00)	(.00)
	40	.98	.96	.98	.96	.98	.98	.98	.98
		(.02)	(.02)	(.02)	(.02)	(.01)	(.01)	(.01)	(.01)
	60	.20	.20	.18	.18	.20	.19	.18	.18
		(.02)	(.02)	(.02)	(.02)	(.02)	(.02)	(.01)	(.01)

Note. Values in () represent SDs per condition; (–) SD was rounded to zero (actual value .004); + = rounded to 1.00.

and 40% of groups were impacted by DIF, increasing the percent of DIF items did not necessarily result in higher power rates. For example, with 20% of DIF items, the estimated power rate was .70, while further increases in the percent of DIF items per group resulted in power rates of .65 and .71, respectively. In conditions with 70% of groups being impacted by DIF, more expected results were noted. Namely, power rates increased as the number of items with DIF increased, from .51, .57, and .63 for the 20%, 40%, and 60%, respectively. Interestingly, when discrimination DIF was .80, power rates slightly decreased as the percent of DIF items increased; this was the case for conditions with 40% and 70% of groups affected by DIF. It should be noted, however, that when discrimination differences were .80, all power rates were very high (ranged from .97 to 1.00).

Panel (b) shows the results for the conditions where DIF was introduced in the item difficulty parameter. It was noted that for all but four conditions, power rates were very high (ranged from .92 to 1.00). However, in conditions where the percent of DIF items was 60%, power rates are unacceptably low (~.20). This result was found regardless of the amount of DIF or the percent of groups that were affected by DIF items. Across all 10-group conditions, standard deviations were very small, suggesting that across the 100 replications within any one condition, power rates were quite homogeneous.

Similar to the 10-group conditions, Type I error rates were zero in all studied conditions in the 20-group cases. Power results for the 20-group conditions are presented in Table 4. Based on these results, there are clear parallels between the 10- and 20-group cases. Similar to the 10-group setting, the non-DIF condition had a Type I error rate

much higher than would normally be expected. In particular, we estimated an operational error rate of 0.18. In other words, 18% of items, on average across replications, were identified as differentially functioning under the condition where items were invariant across groups. Further, and commensurate with the findings for the 10-group case, the Type I error rates were zero to at least the third decimal place under all considered conditions, regardless of DIF location (difficulty or discrimination), magnitude of DIF, percent of items with DIF, or percept of groups with DIF items. Collectively, these findings suggest that the generalized logistic regression method has good *specificity* when some items have DIF and poorer specificity when no items have DIF[e].

Next, Table 4 indicates that, somewhat surprisingly, power to detect DIF items was similar to the 10-group case, suggesting that the number of groups does not matter much for detecting DIF items. Rates of DIF detection were mostly moderate to high, with exceptions similar to the 10-group case. In particular, when 60% of difficulty parameters vary across groups, the power to detect DIF is just .18. And this finding was consistent, regardless of DIF magnitude or the proportion of groups with DIF items. These results suggest that when high proportions of items function differently, the method considered does not do well at detecting these items. According to panel (a), we also found that when the DIF magnitude in the discrimination was small, detection rates were somewhat low (.55 to .72) and that detection rates were the lowest among these (.55 to .68) when a higher percentage of groups (70 compared to 40) had items with DIF in the discrimination parameters. In contrast, the magnitude of DIF in the discrimination was large, detection rates were very high, from .97 when 60% of discriminations had DIF to 1.00 when just 20% of discriminations had DIF. The findings were very consistent across both the 40% and 70% of DIF-group conditions, suggesting that the percentage of groups with DIF items matters little.

With respect to the findings for the 20-group case when DIF was located in the difficulty parameter, Table 4, panel (b) indicates largely very similar findings to the 10-group case. That is, detection rates are high (.96 to 1.00), with the exception of the condition where 60% of items have DIF. And there are slight differences between the 40% and 70% of groups conditions when 40% of items have a small magnitude of DIF. Specifically, the proportion of DIF items detected as having DIF is slightly higher when a lower proportion of groups have DIF (.98 versus .96); however, this is a small difference. And there are no differences in detection rates between the 40% and 70% of groups conditions when 20% of items have DIF in the difficulty parameter, regardless of DIF magnitude. Similar to the 10-group conditions, standard deviations were very small, again indicating relatively homogenous results for power rates across the 100 replications within any one condition.

Discussion and conclusions

As a whole, these findings present a mixed picture with respect to the performance of the generalized logistic regression method when conditions are similar to those found in many ILSA settings. On one hand, it seems that general logistic regression tended to detect DIF well in most cases (see remark on surprising results below). In the presence of DIF, the method was successful in distinguishing between DIF and non-DIF, as evidenced by low Type I error and high power rates. On the other hand, however, in the absence of DIF, the method yielded increased Type I errors.

More specifically, when there are some differentially functioning items in any of the considered conditions, generalized logistic regression does well at ignoring items that are invariant across groups, whether the parameter of interest is the discrimination or difficulty. Further, group size did not appear to be predictive of performance. That is, the findings across the two group sizes (10 and 20) were quite consistent. This finding is in contrast to previous research that used a multiple-groups confirmatory factor analytic approach to detect measurement invariance in large numbers of groups (Rutkowski & Svetina 2014). Specifically, performance of the considered methods depended on the number of groups evaluated, with generally poorer performance connected to larger numbers of groups. Given that international surveys and assessments often feature dozens of system-level participants, a lack dependency on group size is a strength of the method considered here, at least where detection rates were high. And in many cases, power or detection rates were quite high, with more than 90% of DIF items being flagged in 28 of 48 considered conditions. In contrast, power rates were low in 20 of 48 conditions and surprisingly so in 8 of these conditions. Specifically, detecting DIF in discrimination was a challenge for this method when the DIF magnitude was small, with rates ranging from .55 to .72. But perhaps most surprising were the exceedingly low power rates when DIF was located in 60% of difficulty parameters, with rates that ranged from 18% to 20%. One possible explanation for this finding could be that with such a high proportion of DIF items, the method could not accurately identify a set of invariant items against which to compare the non-invariant items – in other words, the presence of DIF items masked the presence of other DIF items; however, it is puzzling that this finding was limited to the difficulty parameter. Nonetheless, uniform DIF is clearly not as easily distinguishable as non-uniform DIF in this particular condition. Lastly, it should be noted that in NONDIF conditions, regardless of the number of groups, Type I error rates were quite inflated. This result is consistent with previous research, which suggested that differences in group means (i.e., impact) may contribute to the inflated Type I error rates (DeMars 2010; Jiang & Stout 1998). As DeMars (2010) suggested, methods such as logistic regression that match groups based on their observed score may not result in well matched examinees on true proficiency. This may lead to false DIF detection due to inaccurate matching[f].

As with any study, our study has limitations due to the study design and the methodology used. With respect to the study design, the generalizability of our conclusions are limited by the selection of manipulated factors in the study. Although we aimed to design our study to closely resemble the context of ILSAs (e.g., rotated booklet design, large and varying sample size, released item parameters), due to space and time constraints, we did not investigate all potential scenarios that ILSAs may occupy in a real context. For example, our focus was only on dichotomously scored data, such that data were generated using a 2-PL IRT model. It is often the case that on large-scale assessments, several item types are present, including items scored dichotomously and polytomously. Further, in the current study, we did not allow for a lower-asymptote parameter, which may be justified for some item types, such as in multiple-choice items. Also, in cases where the effect is rather small (i.e., discrimination or difficulty parameters vary only by a very small amount), a question of the method's performance remains unanswered. These limitations present an opportunity for further research, including a comparison between using the rotated booklet design with planned missing

and conditions where no missingness is present. Additional research should consider this method and its suitability of its use in the context of ILSA studies, as high Type I error rates were observed in the absence of DIF. One avenue of exploration could be to invoke an iterative process called item purification to investigate whether Type I error rates stabilize[g]. Further research should also examine the performance of the studied method in conditions where large amounts of uniform DIF is present, as our results pointed to a rather surprising phenomenon in conditions with 60% of DIF items and difficulty parameter as variant. Namely, power rates in those situations dropped considerably to unacceptable levels, thus further investigations regarding the "breaking point" of the method may be warranted.

From a methodological or practical view, a limitation is that only detection of DIF was investigated in the current study. In practical applications, once DIF is detected in an item or set of items, it is crucial to examine the underlying reasons as to why DIF occurred. As Ferne and Rupp (2007) suggested within the context of language testing, use of expert panels has historically been underutilized, citing reasons of cost and time. Even further, an important consideration by the analyst/testing developers should be made with regards to what is done with items that are flagged as DIF. Although not the focus of the current study, we advocate that detecting DIF in items on assessments is only the first, albeit important, step in testing. Once items are identified as functioning differentially (among the studied groups), substantive analysis should be implemented – for example, use of cultural or linguistic experts may be useful in shedding light on why DIF occurred. With that information, test developers could make more informed decisions to revise and/or remove the flagged items in order to ensure score comparability and appropriate inference regarding groups' performance.

Endnotes

[a]In this paper, we use the term non-DIF to suggest that the item is modeled as having measurement invariance or that no DIF is present in the item (sometimes referred to in the literature as DIF-free).

[b]Preliminary analysis run in *Mplus* (Muthén, & Muthén 1998–2010) for testing configural and metric invariance using similar design choices resulted in a single replication taking almost 24 hours.

[c]The authors studied a total of 27 items; averages and maximums of item parameter differences here were based on 26 items because one item seemed to have provided estimates that were extremely rare (i.e., difficulty parameter estimates between two countries that differed up to 17 points). In order to provide a more general idea of how different item parameters between the countries could be, we eliminated this outlier in our reporting.

[d]This group was chosen as a reference group because it was one of the groups not chosen at random to be impacted by DIF. The selection of this reference group was in that respect at random. This might be slightly different from ILSAs, where the host country of interest may serve as a reference group. Nonetheless, in DIF studies, the choice of reference and/or focal group(s) is often made by the analyst's purpose and goals.

[e]We calculated specificity as 1 – P(item is flagged as DIF | item is non-DIF) or 1 – Type I error rate.

[f]Magis and De Boeck (in press) recently found inflated Type I error rates in Mantel-Haenszel approach, and similarly attribute the inflation to the item impact and difference in item discriminations. Although not the same approach as used in the current study, Magis and De Boeck study found potential sources of Type I error rate inflation as found in DeMars (2010) and the current study.

[g]Item purification is an iterative process, which removes the items currently flagged as DIF from the test scores in order to obtain purified sets of items, unaffected by DIF. Process stops until two consecutive runs yield the same selection of the items. Currently, literature is unclear as to whether item purification controls the Type I error rate to some nominal value and/or maintains higher power (e.g., Magis & Facon 2013).

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
DS and LR reviewed the literature, designed and carried out the analyses, and prepared the manuscript. Both authors read and approved the final manuscript.

Received: 10 February 2014 Accepted: 4 June 2014
Published online: 26 June 2014

References

Baker, FB (1993). *EQUATE2: Computer Program for Equating two Metrics in Item Response Theory [Computer Program]*. Madison: University of Wisconsin, Laboratory of Experimental Design.

Camilli, G., & Shepard, L. A. (1994). *Methods for Identifying Biased Test Items*. Thousand Oaks, CA: SAGE Publications.

DeMars, CE (2010). Type I error inflation for detecting DIF in the presence of impact. *Educational and Psychological Measurement, 70*(6), 961–972. doi:10.1177/0013164410366691.

Dorans, NJ, & Middleton, K (2012). Addressing the extreme assumptions of presumed linkings. *Journal of Educational Measurement, 49*(1), 1–18. doi:10.1111/j.1745-3984.2011.00157.x.

Ellis, BB, & Mead, AD (2000). Assessment of the measurement equivalence of a Spanish translation of the 16PF questionnaire. *Educational and Psychological Measurement, 60*(5), 787–807. doi:10.1177/00131640021970781.

Ercikan, K (2002). Disentangling sources of differential item functioning in Multilanguage assessments. *International Journal of Testing, 2*(3–4), 199–215. doi:10.1080/15305058.2002.9669493.

Ferne, T, & Rupp, AA (2007). A synthesis of 15 years of research on DIF in language testing: methodological advances, challenges, and recommendations. *Language Assessment Quarterly, 4*(2), 113–148. doi:10.1080/15434300701375923.

Fidalgo, AM, & Madeira, JM (2008). Generalized mantel-haenszel methods for differential item functioning detection. *Educational and Psychological Measurement, 68*(6), 940–958. doi:10.1177/0013164408315265.

Fidalgo, ÁM, & Scalon, JD (2010). Using generalized mantel-haenszel statistics to assess DIF among multiple groups. *Journal of Psychoeducational Assessment, 28*(1), 60–69. doi:10.1177/0734282909337302.

Glas, C, & Jehangir, K (2013). Modeling Country-Specific Differential Item Functioning. In L. Rutkowski, M. Von Davier, & D. Rutkowski (Eds.), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis* (pp. 97–116). London: Chapman Hall/CRC Press.

Goodman, J, Wilse, J, Allen, N, & Klaric, J (2011). Identification of differential item functioning in assessment booklet designs with structurally missing data. *Educational and Psychological Measurement, 71*(1), 80–94.

Hambleton, R (2002). Adapting Achievement Tests into Multiple Languages for International Assessments. In A. C. Porter & A. Gamoran (Eds.), *Methodological Advances in Cross-National Surveys of Educational Achievement*. Washinton, DC: National Academies Press.

Hambleton, RK, Swaminathan, H, & Rogers, HJ (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.

Holland, W. P., & Thayer. (1988). Differential Item Performance and the Mantel-Haenszel Procedure. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 129–145). Hillsdale, NJ: LEA.

Jiang, H, & Stout, W (1998). Improved type I error control and reduced estimation bias for DIF detection using SIBTEST. *Journal of Educational and Behavioral Statistics, 23*(4), 291–322. doi:10.2307/1165279.

Jöreskog, KG (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*(4), 409–426. doi:10.1007/BF02291366.

Kim, S-H, Cohen, AS, & Park, T-H (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement, 32*(3), 261–276. doi:10.2307/1435297.

Kulas, JT, Thompson, RC, & Anderson, MG (2011). California psychological inventory dominance scale measurement equivalence: general population normative and Indian, U.K., and U.S. managerial samples. *Educational and Psychological Measurement, 71*(1), 245–257. doi:10.1177/0013164410391580.

Li, Y, Brooks, GP, & Johanson, GA (2012). Item discrimination and type I error in the detection of differential item functioning. *Educational and Psychological Measurement, 72*(5), 847–861. doi:10.1177/0013164411432333.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. New Jersey: Lawrence Erlbaum.

Magis, D, Beland, S, & Raiche, G (2013). *difR: Collection of Methods to Detect Dichotomous Differential Item Functioning (DIF) in Psychometrics. R Package Version 4.4*.

Magis, D., & De Boeck, P. (in press). Type I error inflation in DIF identification with Mantel-Haenszel: An explanation and a solution. *Educational and Psychological Measurement*.

Magis, D, & Facon, B (2013). Item purification does Not always improve DIF detection a counterexample with Angoff's delta plot. *Educational and Psychological Measurement, 73*(2), 293–311. doi:10.1177/0013164412451903.

Magis, D, Raîche, G, Béland, S, & Gérard, P (2011). A generalized logistic regression procedure to detect differential item functioning among multiple groups. *International Journal of Testing, 11*(4), 365–386. doi:10.1080/15305058.2011.602810.

Mellenbergh, GJ (1994). Generalized linear item response theory. *Psychological Bulletin, 115*, 300–307.

Meredith, W (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543. doi:10.1007/BF02294825.

Millsap, RE, & Everson, HT (1993). Methodology review: statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*(4), 297–334. doi:10.1177/014662169301700401.

Muthén, LK, & Muthén, BO (1998–2010). *Mplus user's Guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

OECD. (2010). *TALIS Technical Report*. Paris: OECD Publishing.

OECD. (2012). *PISA 2009 Technical Report*. Paris: OECD. Retrieved from http://www.oecd.org/edu/preschoolandschool/programmeforinternationalstudentassessmentpisa/pisa2009technicalreport.htm.

Oliveri, ME (2012). *Investigation of Within Group Heterogeneity in Measurement Comparability Research (Unpublished PhD Dissertation)*. Vancouver, Canada: University of British Columbia. Retrieved from https://circle.ubc.ca/bitstream/handle/2429/40364/ubc_2012_spring_oliveri_maria.pdf?sequence=3.

Oliveri, ME, & Von Davier, M (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling, 53*, 315–333.

Oliveri, ME, Olson, BF, Ercikan, K, & Zumbo, BD (2012). Methodologies for investigating item- and test-level measurement equivalence in international large-scale assessments. *International Journal of Testing, 12*(3), 203–223. doi:10.1080/15305058.2011.617475.

Oliveri, ME, & Von Davier, M (2014). Toward increasing fairness in score scale calibrations employed in international large-scale assessments. *International Journal of Testing, 14*(1), 1–21. doi:10.1080/15305058.2013.825265.

Olson, J, Martin, MO, & Mullis, IVS (2008). *TIMSS 2007 Technical Report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Partchev, I (2012). *Irtoys: Simple Interface to the Estimation and Plotting of IRT Models. R Package Version 0.1.6*. http://CRAN.R-project.org/package=irtoys.

Penfield, R (2001). Assessing differential item functioning among multiple groups: a comparison of three mantel-haenszel procedures. *Applied Measurement in Education, 14*(3), 235–259. doi:10.1207/S15324818AME1403_3.

Penfield, RD, & Lam, TCM (2000). Assessing differential item functioning in performance assessment: review and recommendations. *Educational Measurement: Issues and Practice, 19*(3), 5–15. doi:10.1111/j.1745-3992.2000.tb00033.x.

Petersen, MA, Groenvold, M, Bjorner, JB, Aaronson, N, Conroy, T, Cull, A, Fayers, P, Hjermstad, M, Sprangers, M, Sullivan, M; European Organisation for Research and Treatment of Cancer Quality of Life Group. (2003). Use of differential item functioning analysis to assess the equivalence of translations of a questionnaire. *Quality of Life Research, 12*(4), 373–385. doi:10.1023/A:1023488915557.

Potenza, MT, & Dorans, NJ (1995). DIF assessment for polytomously scored items: a framework for classification and evaluation. *Applied Psychological Measurement, 19*(1), 23–37. doi:10.1177/014662169501900104.

R Development Core Team. (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

Raju, NS (1988). The area between two item characteristic curves. *Psychometrika, 53*(4), 495–502. doi:10.1007/BF02294403.

Raju, NS (1990). Determining the significance of estimated signed and unsigned areas between Two item response functions. *Applied Psychological Measurement, 14*(2), 197–207. doi:10.1177/014662169001400208.

Rutkowski, L, & Svetina, D (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*(1), 31–57. doi:10.1177/0013164413498257.

Ryan, AM, Horvath, M, Ployhart, RE, Schmitt, N, & Slade, LA (2000). Hypothesizing differential item functioning in global employee opinion surveys. *Personnel Psychology, 53*(3), 531–562. doi:10.1111/j.1744-6570.2000.tb00213.x.

Scott, NW, Fayers, PM, Aaronson, NK, Bottomley, A, De Graeff, A, Groenvold, M, Koller, M, Petersen, MA, Sprangers, MA, EORTC and the Quality of Life Cross Cultural Meta-Analysis Group (2007). The use of differential item functioning analyses to identify cultural differences in responses to the EORTC QLQ-C30. *Quality of Life Research*, 16(1), 115–129. doi:10.1007/s11136-006-9120-1.

Scott, NW, Fayers, PM, Bottomley, A, Aaronson, NK, De Graeff, A, Groenvold, M, Koller, M, Petersen, MA, Sprangers, MA, EORTC and the Quality of Life Cross-Cultural Meta-Analysis Group (2006). Comparing translations of the EORTC QLQ-C30 using differential item functioning analyses. *Quality of Life Research, 15*(6), 1103–1115. doi:10.1007/s11136-006-0040-x.

Stubbe, TC (2011). How do different versions of a test instrument function in a single language? A DIF analysis of the PIRLS 2006 German assessments. *Educational Research and Evaluation, 17*(6), 465–481. doi:10.1080/13803611.2011.630560.

Swaminathan, H, & Rogers, HJ (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361–370. doi:10.2307/1434855.

Thissen, D, Chen, W-H, & Bock, RD (2003). *MULTILOG 7.03*. Chicago: Scientific Software International.

Wilks, SS (1938). Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika, 3*(1), 23–40. doi:10.1007/BF02287917.

Zimowski, MF, Muraki, E, Mislevy, RJ, & Bock, RD (1996). *Bilog-MG: Multiple-Group IRT Analysis and Test Maintenance for Binary Items [Computer Software]*. Chicago: Scientific Software International.

Zumbo, BD (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing, 20*(2), 136–147. doi:10.1191/0265532203lt248oa.

doi:10.1186/s40536-014-0004-5
Cite this article as: Svetina and Rutkowski: **Detecting differential item functioning using generalized logistic regression in the context of large-scale assessments.** *Large-scale Assessments in Education* 2014 **2**:4.